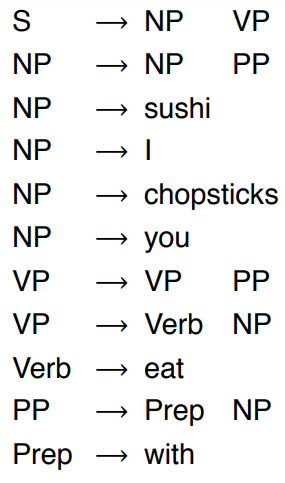
**Syntactic Parsing**

**（一）说明句法解析，对比自顶向下/自底向上、CKY算法、Probabilistic CKY算法、Lexicalization of PCFGs，并说明技术发展脉络和发展动机。**

**答案：**

|  |
| --- |
| 句法解析是自然语言处理中的核心任务之一，旨在分析句子中的语法结构，揭示词语之间的关系。以下是对主要句法解析方法的对比与技术发展脉络的简要说明： 1. ****自顶向下与自底向上解析**** **自顶向下（Top-Down）解析**：从句子的起始符号（通常是S）开始，逐步推导直到生成句子中的各个单词。它采用规则匹配的方式，优先扩展树的根节点。   * + **优点**：易于实现，适用于递归定义的语言结构。   + **缺点**：可能会尝试无效的推导，导致计算量大。   **自底向上（Bottom-Up）解析**：从句子中的词汇开始，逐步构建更高层次的短语结构，直到形成完整的句子树。   * + **优点**：更高效，避免了无效的推导。   + **缺点**：需要更多的语法信息，解析过程较复杂。  2. ****CKY算法**** CKY（Cocke-Younger-Kasami）算法是一种基于动态规划的自底向上解析算法，专为上下文无关文法（CFG）设计，特别适用于Chomsky范式（CNF）。   * **工作原理**：通过构建一个表格，存储部分解析结果，避免重复计算。在表格中，单元格[i, j]保存从第i到第j个词所能构建的所有可能的非终结符。 * **优点**：高效，能在O(n³)时间内完成句子解析。 * **限制**：仅适用于CNF文法，且需要较为精确的语法设计。  3. ****Probabilistic CKY算法**** Probabilistic CKY算法是CKY算法的扩展，结合了概率上下文无关文法（PCFG），为每个规则添加了概率值，解析时选择最可能的树。   * **发展动机**：PCFG的引入使得解析不仅关注语法结构，还考虑了语言的统计性质，有助于提高解析的准确性，尤其是在处理歧义时。 * **优势**：能够选择最可能的解析树，适应语言中的多义性。 * **缺点**：需要大量的标注数据来训练概率模型。  4. ****PCFG的词汇化（Lexicalization of PCFGs）**** 传统的PCFG假设所有非终结符和规则是无条件的，无法捕捉词汇层次的语法信息。词汇化PCFG将词汇信息加入非终结符中，允许模型根据具体词汇选择更合适的语法结构。   * **发展动机**：词汇化PCFG旨在克服传统PCFG的不足，更好地处理词汇对句法结构的影响，提升解析性能，特别是在面对多义词和复杂句式时。 * **优势**：词汇化PCFG能够捕捉到更细粒度的语言特征，提高解析精度。 * **缺点**：计算复杂度增加，需要更大的数据集进行训练。  技术发展脉络  * **初期阶段**：最初的句法解析方法依赖于规则的手工编写，解析效率较低且难以应对语言中的多义性和复杂性。 * **CKY算法的引入**：CKY算法通过动态规划提高了解析效率，使得上下文无关文法的句法分析成为可能。 * **PCFG的出现**：随着统计方法的引入，PCFG为句法分析提供了概率模型，能够处理语言的歧义性和统计规律。 * **词汇化PCFG的发展**：词汇化PCFG的提出进一步提高了句法解析的精度，特别是在处理实际语言中的复杂结构时，充分利用了词汇的信息。  结论 句法解析技术经历了从基于规则的解析到统计方法的转变，随着计算能力和语料库的增加，现代解析方法越来越关注模型的准确性和处理复杂语言现象的能力。自顶向下与自底向上的解析方法各有优劣，CKY算法和其扩展（如Probabilistic CKY算法）在提升效率和准确性方面发挥了重要作用，而词汇化PCFG则进一步推动了句法解析的深度和精度。 |

**（二）给定语法规则：**



1. 请使用CKY算法，写出关于句子“I eat sushi with chopsticks with you”的完整解析过程
2. 请画出最终解析的句法树

**答案：**

|  |
| --- |
| 1.    2. |

**（三）在统计句法分析中，什么是概率上下文无关文法（PCFG）？其如何解决传统语法分析的局限性？**

答案：

|  |
| --- |
| 概率上下文无关文法（PCFG） **定义**： 概率上下文无关文法（PCFG）是上下文无关文法（CFG）的扩展，其中每个语法规则都附加了一个概率值。这些概率表示某个规则在给定语言环境下的生成概率，用于选择最可能的语法结构。 传统语法分析的局限性 传统的上下文无关文法（CFG）仅关注语法结构，假设所有规则在解析过程中是等概率的，忽略了语言中的统计规律和多义性问题。在实际应用中，CFG无法有效处理具有多重可能解析树的句子，例如：   * 句子“Eating fish is good for health”有两种可能的解析：一个是“Eating fish”（名词短语），另一个是“Eating”（动词短语）。  PCFG如何解决这些局限性 **处理歧义性**： PCFG通过为每个语法规则分配概率来处理歧义。当句子具有多种解析方式时，PCFG会通过计算每种解析树的概率，选择最有可能的解析。例如，在上述例子中，PCFG会根据训练数据中出现的频率，为“Eating fish”与“Eating”分配不同的概率，最终选择概率更高的解析树。  **引入统计信息**： 与传统的CFG不同，PCFG利用语料库中的统计数据来为语法规则分配概率。这使得PCFG能够捕捉到语言中的规律性和词汇的依赖关系，例如某些动词在特定句子结构中出现的频率较高。 总结 PCFG通过引入概率，弥补了传统上下文无关文法的不足，解决了语言歧义性和语法规则等概率的假设问题，使得句法分析更符合自然语言的统计特性，从而提高了解析的准确性和鲁棒性。 |

**Semantic Parsing**

**（一）结合语义角色标注（Semantic Role Labeling，SRL）的定义，回答以下问题：**

1. **请说明语义角色标注的主要任务和角色分类标准。**
2. **解释在语义角色标注中，如何使用先验知识（例如选择限制）来解决歧义问题，并举例说明。**

答案：

|  |
| --- |
| 1. 语义角色标注的主要任务和角色分类标准 **任务**： 语义角色标注（SRL）的主要任务是识别句子中的谓词及其相关的角色。例如，在句子“John gave Mary a book”中，SRL的任务是标注“John”为施事（Agent）、“Mary”为接受者（Recipient）、“a book”为主题（Theme）。  **角色分类标准**：   * **施事（Agent）**：执行动作的实体（如“John”）。 * **受事（Patient）**：动作的接受者（如“a book”）。 * **工具（Instrument）**：完成动作所使用的工具（如“a hammer”）。 * **地点（Location）**：动作发生的地点（如“in the park”）。 * **时间（Time）**：动作发生的时间（如“yesterday”）。  2. 使用先验知识解决歧义问题 **选择限制**： 选择限制是指某些动词只能与特定类型的名词搭配，如“eat”通常要求宾语是食物，帮助SRL避免误标。例如，“John ate an apple”和“John ate a book”中，“apple”符合“eat”的选择限制，而“book”则不合适。 |

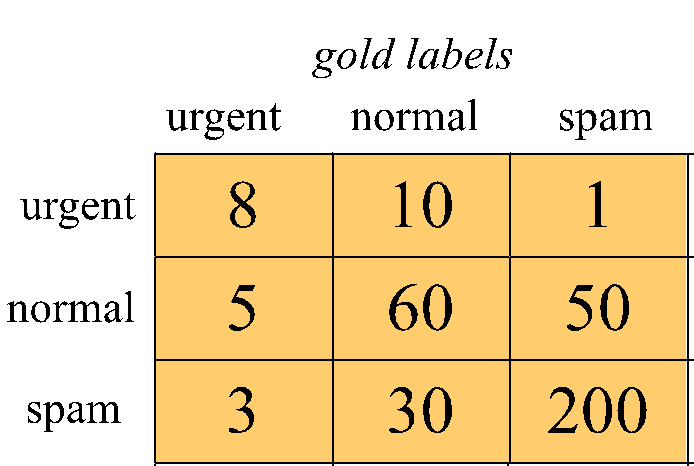
1. PropBank 和 FrameNet 数据集在标注方法上有何不同？请结合各自的具体标注方式（如角色定义与分类）进行说明。

答案：

|  |
| --- |
| **PropBank**：   * **标注方法**：PropBank 主要基于动词及其论元结构进行标注。每个动词都有一组论元（Arg0, Arg1, Arg2等），表示与动词相关的不同角色。 * **角色定义**：例如，动词“buy”可能有：   + **Arg0**（施事）：买东西的人（如“John”）。   + **Arg1**（受事）：被买的物品（如“a book”）。 * **特点**：主要关注动词和它们的论元，标注简单，主要用于句法分析和语义角色识别。   **FrameNet**：   * **标注方法**：FrameNet 基于语境框架（frame）进行标注，框架描述了一个事件或概念及其相关的语义角色。 * **角色定义**：每个框架定义了多个语义角色（例如“JUDGE”和“EVALUATEE”），并通过框架链接多个动词和名词。 例如，在“blame”框架中，“blamer”是施事，“blamee”是受事。 * **特点**：更注重语境和多种语义角色的细化，适用于复杂的事件和情境分析。  总结  * **PropBank** 侧重动词的论元标注，较为简单，适合处理动词及其基本语义角色。 * **FrameNet** 提供了更丰富的语境框架和语义角色，适用于描述复杂事件及其多维度的语义关系。 |

**Text Classification**

（一）给定下面的混淆矩阵



1、请计算出macro-precision，micro-precision

2、请计算出macro-recall，micro-recall

3、请计算出macro-F1，micro-F1

（估算保留4位小数）

答案：

|  |
| --- |
| 1. Precision：     2. Recall：     3. F1：  * （准确值为）  （准确值为） （准确值为） |

### （二）朴素贝叶斯分类器的假设与特点

1、在文本分类中，朴素贝叶斯分类器使用了哪些关键假设？分别解释“条件独立性假设”和“词袋模型假设”。

2、如果训练集中某些类别的文档从未包含特定词汇（如"fantastic"），该如何处理？请简述解决零概率问题的方法。

答案：

|  |
| --- |
| 1. 朴素贝叶斯分类器的关键假设 **条件独立性假设**：假设在给定类别的条件下，文档中的每个词汇是独立的，即一个词的出现与否不受其他词的影响。这个假设简化了计算，使得模型可以独立地计算每个词对类别的贡献。  **词袋模型假设**：假设文档中的词汇顺序不重要，只关心词汇出现的频率。即把文档看作一个无序的词汇集合，每个词的出现次数用于分类。 2. 解决零概率问题的方法 如果某个类别的文档从未包含特定词汇（如“fantastic”），朴素贝叶斯分类器会面临零概率问题。为了解决这个问题，通常使用**拉普拉斯平滑（加1平滑）**。这种方法在计算每个词的概率时，会对词汇的频率加1，从而避免概率为零。这样即使某个词在训练集中没出现，模型也能为其分配一个非零的概率。 |

### （三）文本分类中的偏差与公平性

1. 在情感分析或毒性分类任务中，模型可能对特定群体产生偏见，例如对特定身份（如“女性”或“非裔美国人”）的语言判定为负面情感。这种偏见可能由哪些因素引起？
2. 请结合实际，提出两种缓解模型偏差的策略，并简要说明其实现原理。

**答案：**

|  |
| --- |
| 1. 偏见产生的因素 在情感分析或毒性分类任务中，模型对特定群体产生偏见的原因可能包括：   * **训练数据偏见**：如果训练数据中某些群体的语言使用频率较高，或者负面情感与特定群体相关联，模型可能会学到这种偏差。 * **标注偏差**：如果训练数据中的标注者对某些群体有先入为主的负面看法，这会导致模型学习到不公平的模式。 * **模型架构偏差**：某些模型可能更容易受到特定词汇或特征的影响，从而加剧偏见。  2. 缓解模型偏差的策略 **数据平衡**：通过平衡不同群体的训练数据，使得每个群体的样本数量相对均衡，从而减少模型偏向某一群体的风险。**实现原理**：增加少数群体的数据或减少多数群体的数据，使得模型训练时能够平衡各群体的特征。  **公平性约束**：在训练模型时加入公平性约束，确保模型对不同群体的预测结果相似，避免产生偏见。**实现原理**：通过优化目标函数，限制模型在预测过程中对某一群体的偏差，确保不同群体的结果具有一致性。 |